7-2015

Landmark Classification with Hierarchical Multi-Modal Exemplar Feature

Lei ZHU
Huazhong University of Science and Technology

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

Hai JIN
Huazhong University of Science and Technology

Liang XIE
Huazhong University of Science and Technology

Ran ZHENG
Huazhong University of Science and Technology

DOI: https://doi.org/10.1109/TMM.2015.2431496

Published in IEEE Transactions on Multimedia, 2015 Jul, 17 (7), pp. 981-993.
http://doi.org/10.1109/TMM.2015.2431496

Landmark Classification With Hierarchical Multi-Modal Exemplar Feature

Lei Zhu, Jialie Shen, Hai Jin, *Senior Member, IEEE*, Liang Xie, and Ran Zheng

***Abstract*—Landmark image classification attracts increasing research attention due to its great importance in real applications, ranging from travel guide recommendation to 3-D modelling and visualization of geolocation. While large amount of efforts have been invested, it still remains unsolved by academia and industry. One of the key reasons is the large intra-class variance rooted from the diverse visual appearance of landmark images. Distinguished from most existing methods based on scalable image search, we approach the problem from a new perspective and model landmark classification as multi-modal categorization, which enjoys advantages of low storage overhead and high classification efficiency. Toward this goal, a novel and effective feature representation, called hierarchical multi-modal exemplar (HMME) feature, is proposed to characterize landmark images. In order to compute HMME, training images are first partitioned into the regions with hierarchical grids to generate candidate images and regions. Then, at the stage of exemplar selection, hierarchical discriminative exemplars in multiple modalities are discovered automatically via iterative boosting and latent region label mining. Finally, HMME is generated via a region-based locality-constrained linear coding (RLLC), which effectively encodes semantics of the discovered exemplars into HMME. Meanwhile, dimension reduction is applied to reduce redundant information by projecting the raw HMME into lower-dimensional space. The final HMME enjoys advantages of discriminative and linearly separable. Experimental study has been carried out on real world landmark datasets, and the results demonstrate the superior performance of the proposed approach over several state-of-the-art techniques.**

***Index Terms*—Dimension reduction, diverse visual contents, exemplar selection, hierarchical multi-modal exemplar feature (HMME), landmark classification, region-based locality-constrained linear coding (RLLC).**

Manuscript received April 21, 2014; revised October 21, 2014 and February 13, 2015; accepted April 23, 2015. Date of publication May 08, 2015; date of current version June 13, 2015. This work was supported in part by the National High Technology Research and Development Program of China under Grant 2012AA01A306, and in part by the National Natural Science Foundation of China under Grant 61133008. The work of J. Shen was supported by the Singapore Ministry of Education Academic Research Fund Tier 2 (MOE Ref: MOE2013-T2-2-156). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vasileios Mezaris. *(Corresponding author: Jialie Shen.)*

L. Zhu, H. Jin, and R. Zheng are with the Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China.

J. Shen is with the School of Information Systems, Singapore Management University, 178902 Singapore (e-mail: jlshen@smu.edu.sg).

L. Xie is with School of Science, Wuhan University of Technology, Wuhan 430074, China.

I. Introduction

With the fast advancement of smart mobile devices, high-speed Internet, and photo-sharing Websites, we have witnessed rapid growth of geo-referenced multimedia data in recent years. How to extract the valuable knowledge intelligently from the massive geo-referenced multimedia repositories becomes more and more important [1]–[5].

In particular, as core technical foundation for many location based visual search and analytics applications,[1] effective landmark image classification aims to accurately categorize a query image into a discrete category via learning latent semantics from training images. Due to its great importance, extensive research study has been conducted and consequently many techniques have been proposed in recent years [6]–[10]. However, the problem still remains unsolved and existing methods generally suffer from either low accuracy or poor stability. One of the major reasons for this stagnation is that real images representing landmark categories have highly diverse visual contents. Fig. 1 illustrates a set of good examples about content diversity of images from three representative landmark categories. It is easy to find that diverse visual appearances are commonly caused by three main reasons, listed as follows.

- Landmark consists of a wide range of beautiful sub-regions and sub-components. The images taken at different spots generally have very different visual appearances [as shown in Fig. 1(a)].
- Even for the landmarks from single venue or attraction, diverse visual appearances are caused by photographing the landmarks from various viewpoints [as shown in Fig. 1(b)].
- Visual appearances of landmark images are significantly affected by a wide range of extrinsic factors, such as imaging time, lighting, air quality or weather conditions [as shown in Fig. 1(c)].

Principally, diverse visual appearances inevitably introduce large intra-class visual variance, which poses great challenges on developing accurate landmark classification schemes.

Basic methodologies used in most existing methods can be generally categorized into two independent families: searching based approach [11]–[17] and learning based approach [18], [19]. As the name suggests, basic idea of searching based approach is developed based on scalable image search, where a simple non-parametric k-nearest neighbors (kNN) classifier is applied. The image is predicted as the category which can win the majority votes from k nearest neighbors. As such, order of image rank list has a very strong impact on the final classification performance. In order to achieve high search precision,

[1]Here, landmark is common area of interest at a certain location and there is high likelihood that many people take photos of the area.

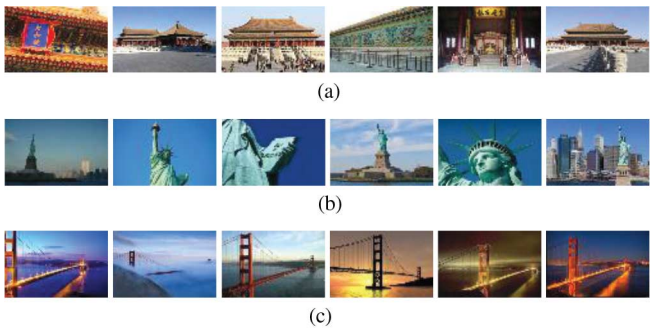

Fig. 1. Images about three representative landmark categories. (a) Images taken at various beauty spots of a landmark. (b) Images taken at various viewpoints. (c) Images taken under various lighting conditions. (a) Forbidden City, Beijing, China. (b) Statue of Liberty, New York City, USA. (c) Golden Gate Bridge, San Francisco, USA.

the image to be classified is firstly used as the query to retrieve a database containing vast amount of images. Then, inappropriate matches over image pairs are removed by post-verification. On the other hand, learning based approach models each landmark as a category and classifies landmark image with discriminative algorithm, which typically has faster classification speed and lower storage costs. Generally, it enjoys three advantages compared to searching based approach: First, with discriminative learning, we only need to store the parameters of classifier models instead of all raw features, which has much lower storage and memory overheads. Second, classification process facilitated by learning based approach is more efficient than that of searching based approach, which is based on time consuming feature matching. Third, discriminative information among different categories can be captured via discriminative learning and redundant information can be removed accordingly. While learning based approach demonstrates good effectiveness, its basic idea is to apply or improve conventional classification models designed for general images, ignoring how to capture characteristics of landmark images [18], [19].

Motivated by the advantages of learning based approach and drawbacks of existing relevant solutions, we approach the problem via modelling landmark classification as a supervised categorization task. Each landmark is treated as a category, and each category has real landmark images with huge appearance differences. In this case, capturing diverse visual contents is essentially important to boost the final categorization performance. Based on the recent literature, two major strategies can be adopted to capture diverse visual contents of landmark images.

- Divide and conquer [20]: Its core idea is to divide the landmark category into sub-categories either automatically or manually, so that images in a sub-category have more visual coherence than difference. For each sub-category, a visual model is trained to characterize the visual distribution. All visual models calculated for sub-categories are integrated together to represent the overall landmark category. At the stage of classification, an image is categorized into a landmark category only if it is categorized into one of its sub-categories. The drawback of this technique is lack of good capability to determine the sub-categories accurately. Imperfect sub-categorization may result in undesirable performance degradation.
- Exemplar-based approach [21], [22]: This approach has been widely applied in visual object detection, where exemplars are representative regions, and similarities between image instance and exemplars are calculated to generate the similarity feature. The advantage of exemplar-based approach is that the noises brought by the inaccurate exemplars can be easily removed by the subsequent machine learning approach. Inspired by this idea, our study explores the idea of exemplar to represent the diverse visual contents of landmark images.

How to compute the signature of landmark images plays an important role in determining the final performance of classification. In this article, we introduce a novel and effective scheme, called hierarchical multi-modal exemplar feature (HMME). To achieve more comprehensive content modelling, candidate images and regions are first generated by partitioning images with hierarchical grids. From these candidate ones, hierarchical exemplars (representative global image views and local regions) in multiple modalities, which represent latent semantics of landmarks, are then discovered via global and regional exemplar selection. Based on the exemplars, HMME is generated by encoding their semantics into different feature dimensions. Dimension reduction is finally conducted via projecting the coarse feature into lower-dimensional space with less redundant information. With the approach, HMME can capture diverse visual contents of landmark images robustly. More importantly, it incorporates heterogeneous discriminative information into a unified feature representation, which enjoys high discriminating capability. The main contributions of this paper can be summarized as follows:

1) An effective exemplar selection approach is proposed to hierarchically discover exemplars in multiple modalities, based on which feature dimensions of HMME are generated.
2) A novel feature generation framework is proposed to encode semantics of the discovered exemplars into HMME, which incorporates heterogeneous discriminative information into a unified feature representation.
3) Comprehensive experiments are conducted on real-world landmark dataset, which includes images with diverse visual contents, to demonstrate the effectiveness of the proposed approach.

The remainder of the paper is structured as follows: We give a detailed review of related work in Section II. Then Section III provides an overview of the HMME based landmark classification system. Section IV presents the details of each part in extraction pipeline of HMME. Next, Section V summarizes the proposed approach and gives time complexity analysis. Experimental configuration is introduced in Section VI. Empirical experimental results and detailed analysis are presented in Section VII. Section VIII finally concludes the paper.

II. Related Work

A. Searching-Based Approach

Most existing approaches for landmark classification are built upon scalable image search [11]–[16]. One of typical examples

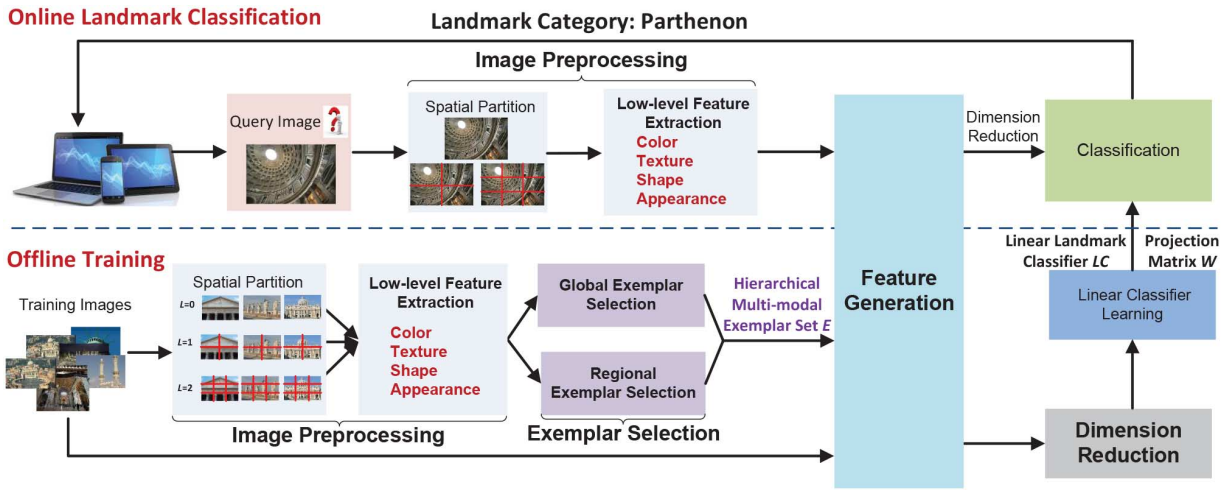

Fig. 2. Framework of the proposed HMME-based landmark classification system.

is that Philbin *et al.* [23] propose to retrieve landmark images represented by bag-of-visual-words (BoVW)[24], where large vocabularies and fast spatial matching are leveraged. In [11], Zheng *et al.* recognize landmark by matching local features of query image against model images with nearest neighbor search. 3D images enjoy better capability to characterize landmarks. Hao *et al.* [14] leverage 3D visual phrase instead of 2D visual phrase to capture spatial structure information and overcome viewpoint changes of landmarks. Chen *et al.* [17] integrate visual content and context for landmark classification. They also propose an approach in [10] to improve the conventional bag-of-visual-phrase (BoVP) [25] for landmark recognition via discriminative learning of category-dependent visual phrases and soft encoding. Nowadays, state-of-the-art landmark search systems are basically built upon variants of BoVW, where the frequency of quantized visual-words is applied as signatures for both query and database images. On the other hand, retrieval response time is very important factor for advanced landmark image search systems. In order to improve search efficiency, index structures, such as inverted file [26], are usually applied to facilitate efficient search. Since searching based approaches are purely based on feature matching, which will inevitably suffer from high storage requirement caused by considerable amount of local features. More importantly, low-level features may fail to capture diverse visual contents of landmark images.

The landmark classification techniques discussed above are designed for desktop environment. With recent popularity of mobile search, many different approaches have been recently designed specifically for mobile platform [6], [7], [10], [13], [27], [28]. However, their main research focus is on how to design intelligent algorithm to calculate compact descriptors to reduce memory consumption and improve network transmission. Due to the limited space here, we won't detail them.

B. Learning-Based Approach

Comparing to searching based approach, less schemes are proposed to classify landmark based on discriminative learning [18], [19]. Li *et al.* [18] learn a landmark classifier, which combines heterogeneous information from visual contents and textural tags in framework of support vector machine (SVM) [29]. Visual contents are characterized by vector quantized local features, while textural information is described by a frequency vector whose dimensions denote frequently used tags. Image and text feature vectors are normalized and simply concatenated into a unified feature. Bergamo *et al.* [19] propose to leverage structure from motion to learn discriminative codebooks for local features, which are specially designed for BoVW based classification model. There is an effective approach [30] that need to be noted here for its superior performance on task of general image classification. In [30], a novel combination scheme is proposed to integrate the multiple features extracted from different visual modalities with optimal combination weights via multiple kernel learning (MKL)[31]. Experimental results show that the combined feature performs better than any single feature. It can be considered as one of the most effective multiple feature fusion approaches.

III. System Overview

In this section, we briefly introduce the proposed HMME based landmark classification system. Fig. 2 illustrates its basic architecture of the framework. It is mainly comprised of two major components: offline training and online classification.

During the offline training, at the stage of image preprocessing, training images are first partitioned into multiple regions with hierarchical standard grids. Four visual features from heterogeneous visual modalities are extracted from each region to represent visual contents. Next, at the stage of exemplar selection, hierarchical exemplars are automatically discovered in multiple modalities. After that, at the stage of feature generation, optimal reconstruction coefficients, between training images and the discovered exemplars, are learned from multiple visual modalities and spatial levels, and combined to construct the feature dimensions of HMME. Finally, at the stage of dimension reduction, dimension projection matrix W is learned directly on training features. With W, the features of all the images in database are projected into lower-dimensional space. Hierarchical multi-modal exemplar set E, landmark

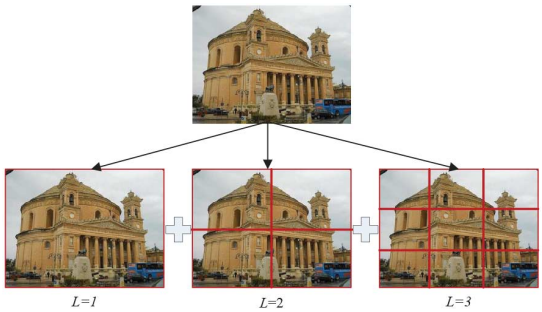

Fig. 3. Image is hierarchically partitioned into regions without overlapping.

classifier LC, and projection matrix W are preserved for online classification.

In the online landmark classification, image preprocessing is first performed on query image. Next, HMME is extracted based on E, and projected into lower-dimensional feature space as training images with W. Then, the projected query feature is imported into the pre-learned LC. Finally, landmark category of query image is obtained and returned back to user.

IV. Hierarchical Multi-Modal Exemplar Feature Extraction

In this section, we give a detail introduction of core components in computing HMME.

A. Image Preprocessing

Image preprocessing aims to extract visual features of landmark images. Each image is first partitioned into regions without overlapping by using hierarchical standard grids (as shown in Fig. 3). Then, four widely used low-level visual features are extracted from different regions. Details of them are as follows:

- *Color Moments (CM)* [32]. Image is first partitioned into regions without overlapping with 3×3 grid. In each segmented region, color mean, color variance, and color skewness are then extracted for each color channel in HSV color space. Features calculated from regions are finally concatenated to form 81-dimensional vector.
- *Local Binary Pattern (LBP)* [33]. LBP is simple yet powerful texture descriptor to describe local structure of pixel by comparing centroid pixel with surrounding pixels. It has good property of tolerating regarding illumination changes. In this study, 58-dimensional LBP is adopted for texture description.
- *Histogram of Oriented Gradients (HOG)* [34]. HOG counts occurrences of gradient orientation in localized portions of an image. The feature dimension of HOG used in this study is 31.
- *Bag-of-Visual-Words (BoVW)* [35]. BoVW quantizes order-less local features to visual-words and represents image as frequency histograms of visual-words. Densely sampling strategy is employed to detect interest points and scale invariant feature transform (SIFT) [36] is used to describe image patches. Each interest point is represented by a vector of 128 dimensions.

B. Exemplar Selection

One of the potential reasons why landmarks are so fascinating and attractive is that each landmark owns unique peculiarities. It is common that, when touring a landmark, the same spectacular scenery spots will be frequently photographed by many different tourists unhesitatingly. In contrast, there are not many tourists who take photos about uninteresting and unattractive places or venue. Consequently, the recorded landmark images often illustrate multiple views of landmarks. They could be distinctive global image views and local regions, which attract tourist's attentions. From viewpoint of visual representation, diverse visual contents about landmark categories can be represented on bases of these representative image views and regions, or in other words, exemplars. Thus, effective exemplar modelling from multiple spatial levels for landmark visual representation is important. A compact and discriminative exemplar set is required to achieve effective landmark image representation. In light of the observation, we propose discriminative exemplar selection approach to discover informative exemplar hierarchically.

1) Global Exemplar Selection: Global exemplar selection is one of the most important system components whose main functionality is to discover representative global image views for each landmark category. To achieve the goal, we propose boosting-based global exemplar selection (BGES). The main idea of BGES is to exploit classification errors of the learned weak classifiers to measure the discriminating capability of images, and then select the most discriminative ones. we manually label several images and adopt them as training image set. Each image in training set is regarded as candidate exemplar, which is chosen as the only positive image. From the images belonging to different categories in training image set, we select several images that are visually similar to positive image, and adopt them as the negative images. In this study, image similarity is measured with the feature in corresponding modality with Euclidean distance. Weak classifier is trained to separate positive image from the negative images by a large margin. At each iteration, classification errors of all weak classifiers are calculated by comparing the predicted labels with true labels of training images. Weak classifier with the minimum classification error is chosen as the current most discriminative classifier, and its corresponding image is considered as the current most discriminative exemplar. With this measurement, at each iteration, one exemplar is removed from candidate exemplar set and added into global exemplar set. According to the basic idea of boosting, the iterative process will automatically stop when the minimum classification error is above 0.5. This setting is reasonable as weak classifiers trained in this case are not discriminative enough. When fine-tuning BGES system parameters, we find two important aspects.

- Besides exemplar selection, there is a weight adjustment process in each iteration, which readjusts the importance of images according to classification errors. Via this procedure, weak classifiers gradually focus their main attentions on separating images that still cannot be distinguished by previously selected weak classifiers. In this way, weight adjustment process can be considered as a hidden and au-

tomatic sub-category grouping. Therefore, the image corresponding to the selected weak classifier in each iteration can be considered as the representative image in sub-category, and the current most discriminative global exemplar accordingly.

- Weaker classifiers explored in BGES are actually sub-category classifiers. Weak classifiers provide a way of measuring discriminative capability for candidate exemplars. Via discriminative learning, the global exemplars are embedded with semantics of sub-categories.

Let N denote the number of training images. $I = \{I_1, \ldots, I_N\}$ denotes training image set. As shown above, candidate exemplar set serves as actual training image set. To avoid confusion, they are denoted using different symbols. Let $\{I_i\}_{i=1}^{S}$ and $\{I_j\}_{j=1}^{C}$ denote candidate exemplar set and training image set respectively. $\{(g_i^p, z_i)\}_{i=1}^{S}$ and $\{(f_j^p, y_j)\}_{j=1}^{C}$ denote their feature representations in modality p, where $p = 1, 2, \ldots, P$ and P is the number of modalities. S denotes the number of training images, C denotes the number of candidate exemplars, $S = C = N$. g_i^p, f_j^p denote d^p dimensional features extracted from exemplar image I_i and training image I_j respectively. $z_i, y_j = 1, -1$ denote that image is positively labeled and negatively labeled respectively. The main aim of BGES is to construct global exemplar set E_1.

In BGES, we develop exemplar-based weak classifier based on SVM due to its high classification rate and time efficiency. For implementation, we use LIBSVM [37] as base to build weak classifiers. For a candidate exemplar I_j, we train a weak classifier H_j^p in modality p to separate I_j to negative images by a large margin. Let us denote $\Theta_j^p = \{(\omega_j^p, b_j^p)\}$ as parameters of H_j^p, ω_j^p and b_j^p are soft margin parameter and bias multiplier respectively. These parameters are calculated by solving the following optimization problem:

$$\arg\min_{\Theta_j^p} \|\omega j^p\|^2 + \alpha \sum_{i=1}^{S} \max\{0, 1 - z_i(\omega_j^p \cdot \psi(g_i^p) + b_j^p)\} + \beta \max\{0, 1 - y_j(\omega_j^p \cdot \psi(f_j^p) + b_j^p)\} \quad (1)$$

where $\alpha, \beta > 0$ are regularization parameters, which play a trade-off between marginal separation and error penalty, ψ is hidden function which maps the linearly inseparable low-level features into high-dimensional and linearly separable features. In this work, we simply use explicit map proposed in [38] to finish this step. With the learned parameters, weak classifier of I_j in modality p can be constructed as

$$H_j^p(g_i^p) = \mathrm{sgn}(\omega_j^p \cdot \psi(g_i^p) + b_j^p). \quad (2)$$

At each iteration, we calculate total error of weak classifier by summing its prediction errors on all training images

$$e_j^p = \sum_{i=1}^{S} w_{t,i} |H_j^p(g_i^p) - z_i| \quad (3)$$

where $w_{t,i}$ is weight of training image I_i calculated at iteration t. We choose weak classifier with the lowest error rate at tth iteration

$$H_{min\_idx}^p = \arg\min_{H_j^p} e_j^p, \quad min\_idx \in \{1, 2 \ldots, C\}. \quad (4)$$

After that, the global image I_{min_idx} corresponding to $H_{min\_idx}^p$ is added into the global exemplar set in modality p.

$$E_1^p = E_1^p \cup I_{min\_idx} \quad (5)$$

Algorithm 1: BGES in modality p

Input:
Training image set $\{(g_i^p, z_i)_{i=1}^{S}\}$.
Candidate exemplar set $\{(f_j^p, y_j)_{j=1}^{C}\}$.
Output:
Global exemplar set in modality p, E_1^p.
Initialize weights of training images with $1/S$, $count = 0$.
while $count < T_1$ **do**
  Normalize weights of training images.
  for I_j in candidate image set **do**
    Prepare positive image and negative images.
    Train weak classifier H_j^p via (1).
    Calculate classification errors of H_j^p via (3).
  end for
  Choose the weak classifier $H_{min\_idx}^p$ via (4).
  if $e_{min\_idx}^p > 0.5$ **then** break.
  Add I_{min_idx} that corresponds to $H_{min\_idx}^p$ into E_1^p, and simultaneously remove I_{min_idx} from candidate exemplar set. Let $E_1^p = E_1^p \cup I_{min\_idx}$, $\epsilon_t = e_{min\_idx}^p$, and $count = count + 1$.
  Update weights of training images: $w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$, where $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$, and $e_i = 0, 1$ for incorrect and correct classification respectively.
end while

Algorithm 1 shows the detailed procedure of BGES for modality p. We perform BGES in each modality, obtaining exemplar set $\{E_1^p\}_{p=1}^{P}$ and count the occurrence frequency of each exemplar image in $\{E_1^p\}_{p=1}^{P}$. Formally, the occurrence frequency of exemplar image I_j, $Freq(I_j)$ can be defined as follows:

$$Freq(I_j) = \sum_{p=1}^{P} \delta(I_j, E_1^p), \quad \delta(I_j, E_1^p) = \begin{cases} 1 & I_j \in E_1^p \\ 0 & I_j \notin E_1^p \end{cases}. \quad (6)$$

The occurrence frequencies of exemplar images are ranked in descending order. The first T_1 exemplars are added to the global exemplar set $E_1 = \{E_{1t}\}_{t=1}^{T_1}$.

2) Regional Exemplar Selection: Regional exemplar selection is to discover representative local regions for each landmark. Principally, a desirable regional exemplar set should at least have three properties:

- *Relevance*. This property aims to select regional exemplars that are most probably relevant to the landmark. This is because that many irrelevant regions, such as regions that describe the background sky or grass, are generated due to spatial partition.
- *Discrimination*. This property aims to select regional exemplars that are most discriminative for classifier training.

This property guarantees that less unused and redundant regions are included in regional exemplar set.

- *Representativeness*. This property aims to select regional exemplars that can represent the underlying visual distribution of regions. Representative regions from dense regions are enough for image representation and excessive regions expand regional exemplar set.

In our scheme, image is partitioned into regions with $l \times l$ grid at spatial level l. Regions are labelled sequentially as 1 to l^2 from left to right and top to bottom. $I_i = \{I_{ij}\}_{j=1}^{l^2}$. The size of region I_{ij} is $\frac{1}{l^2}$ of size of global image I_i. We redefine the training image set in modality p as $\{(f_{ij}^p, y_i)\}, i = 1, \ldots, N, j = 1, \ldots, l^2$, f_{ij}^p denotes d^p dimensional low-level features in modality p which is extracted from jth region of image I_i, y_i is image label. In this case, all the images in training image set are used to discover discriminative regional exemplars at spatial level l, $E_l = \{E_{lt}\}_{t=1}^{T_l}$.

To achieve effective regional exemplar selection, we formulate the process based on multi-modal multi-instance learning (M^3IL). Thus image and region are represented as "bag" and as "instance" respectively under the framework. Distinguished from conventional multiple instance learning (MIL) [39] which learns a unified discriminative set classifier, we extend MIL in multiple modalities to discover discriminative regional exemplars. In our setting, M^3IL is formulated as a mixed integer programming problem, which simultaneously calculates optimal latent category labels of regions, and finds the optimal separation hyperplane that maximizes the separation margin. Its formulation is as follows:

$$
\begin{aligned}
&\min_{\{y_{ij}\}} \min_{w^p, b^p} \sum_{p=1}^{P} \|w^p\|^2 + \eta \sum_{p=1}^{P} \|y_{ij}^p - y_{ij}\|^2 + \gamma \sum_{p=1}^{P} \sum_{i=1}^{N} \sum_{j=1}^{l^2} \xi_{ij}^p \\
&s.t. \quad y_{ij}^p(< w^p \cdot \psi(f_{ij}^p) > + b^p) \geq 1 - \xi_{ij}^p, \xi_{ij}^p \geq 0, \ \forall p, \forall i, \forall j \\
&\quad \sum_{j=1}^{l^2} y_{ij}^p \geq 1, \sum_{j=1}^{l^2} y_{ij} \geq 1, if \ y_i = 1, \quad \forall p \\
&\quad y_{ij}^p = -1, y_{ij} = -1, if \ y_i = -1, \quad \forall p
\end{aligned} \tag{7}
$$

where $\eta, \gamma > 0$ are regularization parameters which play trade-off between three terms. w^p and b^p are soft margin and bias multiplier parameter respectively. y_{ij}^p is the estimated label of region I_{ij} by using single feature in modality p. y_{ij} is the label of region I_{ij} estimated by all the features. The term $\sum_{p=1}^{P} \|y_{ij}^p - y_{ij}\|^2$ is to guarantee that regional labels obtained by different modalities should be consistent. It should be noted that, when $\eta = 0$, the above optimization problem is degenerated to multiple MIL learners in P modalities. In implementation, this problem is solved with alternate optimization. For given hidden labels, we develop the optimization formulation by using standard LIBSVM solution. On the other hand, for a given discriminative model, we minimize the objective function by updating the hidden labels. Algorithm 2 shows details about the region exemplar selection procedure.

Algorithm 2: M^3IL based regional exemplar selection

Input:
Training image set $\{(f_{ij}^p, y_i)\}$,$i=1, \ldots, N, j=1, \ldots, l^2$.
Output:
Latent labels of regions $\{y_{ij}\}$,$i=1, 2, \ldots, N, j=1, \ldots, l^2$.

```
for i = 1 : N, j = 1 : l^2 do
    Initialize y_ij = y_i.
end for
repeat
    for p = 1 : P do
        Train SVM with regional feature f_ij^p and label y_ij^p.
        for each y_i = 1 do
            Compute classification scores of regions via
            F_ij^p = sgn(< w^p · ψ(f_ij^p) > + b^p).
        end for
    end for
    for each y_i = 1 do
        y_ij = sgn(Σ_{p=1}^{P} F_ij^p), j = 1, 2, ..., l^2.
        If Σ_{j=1}^{l^2} y_ij = 0 then
            j' = arg max_{j∈{1,2,...,l^2}} F_ij, y_ij' = 1
        end if
    end for
until y_ij has not been changed
```

After solving M^3IL, possible labels of all regions are obtained directly. Specifically, $y_{ij} = 1$ and $y_{ij} = -1$ mean that the corresponding region is relevant and irrelevant to the landmark respectively. From the relevant regions, we further select the most representative ones. We first concatenate four visual features extracted from region to represent it. Then, k-means clustering [40] is applied to the relevant regions. The regions which have the nearest distance with clustering centres are considered as representative regional exemplars (each cluster centre corresponds to one region). This strategy is reasonable as these regions can also represent regions in their clusters as centres, which matches the requirement of regional exemplar. It should be noted that k-means used in this paper can be substituted by any other effective clustering algorithms. Selecting the most effective clustering algorithm is out of the main scope of this study. T_l exemplars $E_l = \{E_{lt}\}_{t=1}^{T_l}$ are selected as regional exemplars at spatial level l.

C. Feature Generation

We perform global and regional exemplar selection at each spatial level, obtaining hierarchical multi-modal exemplars $E = \{E_l\}_{l=1}^{L} = \{E_{lt}, l = 1, \ldots, L, t = 1, \ldots, T_l\}$. The cardinality of exemplar set is $\sum_{l=1}^{L} T_l$.

Landmark image can be described from various perspectives. Certain types of landmark images can be comprehensively characterized by specific low-level features. Descriptive information in features from heterogeneous modalities are complementary with each other and the effective combination can boost overall performance [17], [18], [30]. Therefore, we concatenate reconstruction coefficients calculated from different modalities to develop the feature dimensions of HMME.

With the image I's HMME feature $f(I)$, it can be derived as

$$f(I) = [(f^1(I))^{\mathrm{T}}, \ldots, (f^p(I))^{\mathrm{T}}, \ldots, (f^P(I))^{\mathrm{T}}]^{\mathrm{T}}. \quad (8)$$

The dimension of $f(I)$ is $P \times \sum_{l=1}^{L}(l^2 \times T_l)$. $f^p(I)$ is sub-feature of HMME computed in modality p, its computational formula is

$$f^p(I) = [(f_1^p(I))^{\mathrm{T}}, \ldots, (f_l^p(I))^{\mathrm{T}}, \ldots, (f_L^p(I))^{\mathrm{T}}]^{\mathrm{T}}. \quad (9)$$

The dimension of $f^p(I)$ is $\sum_{l=1}^{L}(l^2 \times T_l)$. $f_l^p(I)$ is sub-feature of HMME computed in modality p and at spatial level l.

To generate the sub-feature of HMME to describe the visual contents of the segmented region, we propose a region-based locality-constrained linear coding (RLLC). Its basic idea is to calculate the optimal reconstruction coefficients between image instance and the discovered exemplars, so as to encode more semantics of exemplars into HMME. Note that, feature vector generated in this way enjoys desirable advantage of linearly separable.

For database image I_i, $V_{ij}^p \in \mathbb{R}^{T_l \times 1}$ is the sub-feature of HMME extracted on region I_{ij} and in modality p. Thus, V_{ij}^p can be computed by solving the following optimization problem:

$$\min_{V_{ij}^p} \sum_{p=1}^{P} \sum_{i=1}^{N} \sum_{j=1}^{l^2} \|f_{ij}^p - U_l^p V_{ij}^p\| \quad s.t. \quad \mathbf{1}^T V_{ij}^p = 1, \forall p, \forall i, \forall j \quad (10)$$

where $U_l^p = [U_{l1}^p, U_{l2}^p, \ldots, U_{lT_l}^p]$ denotes low-level features of regional exemplars in modality p and at spatial level l. Its feature dimension is $d^p \times T_l$.

After solving the above optimization problem in (10), we obtain sub-features of image I_i at spatial level l

$$f_l^p(I_i) = [(V_{i1}^p)^{\mathrm{T}}, \ldots, (V_{il^2}^p)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{(l^2 \times T_l) \times 1}. \quad (11)$$

For a query image $I_q = \{I_{q1}, I_{q2}, \ldots, I_{ql^2}\}$, we calculate the sub-feature of HMME on region I_{qr} by solving the following optimization problem:

$$\min_{V_{qr}^p} \sum_{p=1}^{P} \|f_{qr}^p - U_l^p V_{qr}^p\| \quad s.t. \quad \mathbf{1}^T V_{qr}^p = 1. \quad (12)$$

HMME of query image I_q at spatial level l can be represented as

$$f_l^p(I_q) = [(V_{q1}^p)^{\mathrm{T}}, \ldots, (V_{ql^2}^p)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{l^2 \times T_l}. \quad (13)$$

D. Dimension Reduction

There are mainly two types of information redundancy in raw HMME: (1) redundant information among multiple modalities. Feature dimensions generated by different modalities may have the equivalent discriminating ability. In principle, these "overlapping" dimensions can be compressed directly without accuracy loss. (2) redundant information among multiple spatial levels. Discriminative information extracted from several spatial levels may be enough on distinguishing the images. Hence, concatenating features from excessive spatial levels may cause redundancy.

To remove redundancy, our approach applies dimension reduction (PCA [41]) as postprocessing to further project the coarse HMME into lower-dimensional space. Dimension reduction has an additional positive effect that the reduced feature dimensions can speedup the subsequent processing. With learning on training features, the projection matrix W is obtained. HMMEs of all training and testing images are projected into lower-dimensional space with W. After projection, landmark classifier LC is trained and reserved for online classification.

Algorithm 3: HMME based landmark classification

Input:
Query image and training images.
Output:
Landmark category of query image.
Offline Training
1: Image preprocessing as shown in Section IV-A.
2: Discover hierarchical multi-modal exemplar set E via BGES and M^{3}IL as shown in Section IV-B.
3: Generate HMMEs for training images as shown in (8).
4: Learn projection matrix W with PCA as shown in Section IV-D.
5: Project HMMEs into lower-dimensional space with W.
6: Train landmark classifier LC on the projected features with Linear SVM [42].
7: Output E, W, and LC, for online classification.
Online Landmark Classification
8: Image preprocessing as shown in Section IV-A.
9: Extract HMME for query image based on E as shown in (8).
10: Project query feature with W.
11: Import the projected query feature into LC and obtain the estimated landmark category.

V. Summary

This section summarizes the proposed approach and gives a comprehensive computation complexity analysis. Algorithm 3 describes the complete algorithmic steps of HMME based landmark classification system. Offline training is comprised of 7 main steps, while online landmark classification is comprised of 4 main steps. At the stage of offline training, the process of spatial partition (step 1) can be finished in $O(N)$ as there are N training images. Assuming that the selection for one exemplar can be completed in $O(1)$, the computation cost of exemplar selection (step 2) is $O(\sum_{l=1}^{L} T_l)$, as there are $\sum_{l=1}^{L} T_l$ discriminative exemplars. Since there are P different modalities and N training images, computations needed for accomplishing the process of feature generation for all training images (step 3) is $O(N \times P \times \sum_{l=1}^{L} T_l)$ (region coding is assumed to be completed in $O(1)$). Denote R as the dimension of the reduced feature, time complexity of dimension reduction (step 5) is $O(N \times R)$. The computation complexity of landmark classifier training (linear classifier training) is $O(N)$ (step 6). At the stage of online landmark classification, the process of feature extraction and dimen-

Fig. 4. Typical images sampled from *Landmark-25* and *Landmark-101*. Our collected landmark images that are photographed from different viewpoints, under different lighting conditions, and for different beauty spots.

sion reduction for a given query image (step 8 and step 9) can be completed in $O(P \times \sum_{l=1}^{L} T_l)$ and $O(R)$, respectively. Computation complexity of online classification (step 10, linear classification) are $O(1)$.

VI. Experimental Configuration

A. Experimental Datasets

We develop two landmark datasets containing worldwide landmarks distributed throughout the earth by crawling images from Flickr.[2] For a specific landmark, candidate images are first obtained by retrieving images from Flickr with relevant keywords and the provided API. Due to low accuracy of text-based image search, candidate landmark images contain many outliers and low-quality images. We manually remove undesirable images via the procedure as following: First, the irrelevant images are excluded and then artificially processed images, such as images with black wire frame, are excluded to retain raw information of images. After that, we remove low-quality images, such as images that suffer from severe motion blur or overexposure. Finally, images with human faces are excluded to avoid privacy breaches. From the processed results, we collect images to construct two landmark datasets. The first landmark dataset we collected has 25 landmark categories, while the second one holds 101 landmark categories. It should be noted that both two datasets include the images photographed for various beauty spots, from various viewpoints, and under various weather conditions. Therefore, these two datasets are challenging for classification as the visual appearance of images in a landmark category is more diverse. We denote them as *Landmark-25* and *Landmark-101* in experiment. Typical images from them are shown in Fig. 4.

In addition, we also conduct experiments on a publicly available landmark dataset, *Landmark-3D*.[3] This dataset is developed by Hao *et al.* in [14], which includes 45 K images in 25 landmark categories. It has also been used in recent literature [19]. All the images also come from Flickr with manual filtering. This dataset mainly contains images that are photographed for landmark of single construction, images in a landmark category have less visual diversity.

Similar to testing method in [18], for three datasets, 200 images from each category are randomly selected to comprise the testing dataset, making classification results easier to interpret. 100 images are used for training and the remaining images are used for performance evaluation.

B. Compared Approaches

We conduct experiment to compare the performance of the approach against the state-of-the-art techniques. Details of the approaches used for comparison are as follows.

1) $MLF + kNN$[12]: In this approach, multiple low-level features (MLF) are combined with kNN classifer. kNN is a typical searching based approach. It estimates the category of landmark via a pure data-driven scene matching. The best performance of $MLF + kNN$ is achieved when k is set to 20, and the feature combination weight (CM, LBP, HOG, BoVW) is set to 0.1, 0.3, 0.2, 0.4.
2) $MLF + SVM$[18]: This approach adopts multiple low-level features (MLF) as image descriptor and SVM as classifier. *MLF+SVM* is a typical learning based approach, where all employed features are concatenated into a single feature vector and SVM is employed as the underlying classifier. In implementation, SVM with Gaussian kernel is used to measure similarities of images. Smoothing factor and cost parameter are set to 1 and 10 respectively to maximize the performance.
3) $MLF + MKL$[30]: This approach adopts multiple low-level features (MLF) as image descriptor and MKL as classifier. Different from $MLF + SVM$, MKL can automatically weight the importance of each feature according to their discriminating ability. The best performance of $MLF + MKL$ is achieved when the feature combination weight (CM, LBP, HOG, BoVW) is set to 0.1, 0.4, 0.1, 0.4.

Two settings of our approach are tested:

1) $HMME+kNN$: This approach applies HMME as visual descriptor and kNN as classifier. Our approach in this setting is used to demonstrate the performance when HMME is combined with non-parametric classifier. The best performance of $HMME + kNN$ is achieved when k is set to 20.
2) $HMME + SVM$: This approach uses HMME as visual descriptor and SVM as classifier. Since HMME is linearly separable, linear kernel is used to measure the similarities between images. Regularization parameters α, β in (1) and η, γ in (7) are tuned using $\{10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}\}$. We obtain the experimental results when $\alpha = 0.1, \beta = 0.01, \eta = \gamma = 10$.

C. Implementation Details

All the training images are considered as candidate image set. Therefore, in global exemplar selection, the size of candidate exemplar set and training image set is set to 2500 on *Landmark-25* and *Landmark-3D*, and that is 10100 on *Landmark-101*. All the images are hierarchically partitioned into re-

[2][Online]. Available: http://www.flickr.com

[3][Online]. Available: http://landmark3d.codeplex.com/

TABLE I
CLASSIFICATION ACCURACIES ACHIEVED BY DIFFERENT APPROACHES ON LANDMARK DATASETS. THE ITEMS SHOWN IN BOLD ARE THE TWO BEST RESULTS. THE RESULTS WITH THE ASTERISK ARE THE BEST

Landmark datasets	Approaches (%)				
	MLF+kNN [12]	*MLF+SVM* [18]	*MLF+MKL* [30]	*HMME+kNN*	*HMME+SVM*
Landmark-3D	67.20±0.90	82.7±0.60	89.8 ±0.76	**97.76±0.54**	**99.84±0.65***
Landmark-25	33.68±0.59	42.04±0.97	54.96±0.88	**67.88±0.63**	**73.72±0.49***
Landmark-101	27.08±0.38	38.42±0.40	54.46±0.52	**58.51±0.44**	**64.52±0.58***

gions from level 1 to 3 with 1×1, 2×2, and 3×3 grids, generating 1, 4, and 9 regions respectively without overlapping. Each region is represented by four visual features (as described in Section IV-A). In order to extract BoVW, the best visual vocabulary is 100 on *Landmark-25* and *Landmark-3D*, and that is 1000 on *Landmark-101*. The best initial feature dimension of HMME on *Landmark-25* and *Landmark-3D* is 23000, and that is 92920 on *Landmark-101*. The best reduced size of HMME on *Landmark-25* and *Landmark-3D* is 2200, and that is 10000 on *Landmark-101*. In regional exemplar selection, the number of negative images for constructing weak classifiers is set to 3 to maximize performance. For SVM implementation, we use LIBSVM (C-SVC)[37] to train classifiers, which solves landmark classification as a multi-class problem.

D. Evaluation Metrics

Experimental performance of different approaches are evaluated on standard metric: classification accuracy. It is defined as the ratio of number of correctly classied test images to the total number of test images. A test image is considered to be correctly classied only if the estimated landmark category label of query image matches with the ground-truth category label. All our experiments have been run on the platform equipped with an Intel Core i7 920 CPU running at 2.67 GHz. The operating system is 64-bit RHEL AS 5.4 with Linux kernel 2.6.18. All the experiments in this study are performed 10 times on randomly selected training and testing images. And we also report final classification rates with standard deviation.

VII. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we first present performance comparison results, and then provide discussions on factors that have the greatest impact on the performance.

A. Performance Comparison Results

Table I summarizes main experimental results. We can easily find from the table that, our proposed approaches outperform their competitors on all datasets. For example, on *landmark-25*, $HMME+kNN$ and $HMME+SVM$ achieve classification accuracy 67.88% and 73.72% respectively, which are 13% and 19% higher compared with the classification accuracy achieved by the second best of approach, $MLF + MKL$. In addition, a few important findings can be found as follows.

- The approaches using SVM as classifier generally perform better than the approaches that adopt kNN as classifier e.g. $HMME+SVM$ is better than $HMME+kNN$, $MLF + SVM$ is better than $MLF+kNN$. This is because that SVM has better discriminative capability than kNN when separating high-dimensional visual features. However, this is a special case in experiment that, $HMME+kNN$ even achieves better performance than $MLF + MKL$, where MKL is used as classifier. This is because, HMME is linearly separable, which can be separated well with simple classifier kNN. This experimental result demonstrates the advantage of HMME that it can involve rich discriminative information into feature representation, so that simpler classifier can be adopted for good performance.
- Compared with $MLF + SVM$, $MLF + MKL$ performs better. The reason is that MKL learns appropriate combination weights according to the discriminating ability of features on specific landmark category. The poor performance of *MLF+SVM* observed in this experiment also verifies that effective landmark classification cannot be achieved by directly concatenating low-level features.
- Although *landmark-3D* and *landmark-25* include the same number of landmark categories, classification accuracies obtained on *landmark-25* are all lower than that achieved on *landmark-3D*. The performance gap is more than 20%. Classification task on *landmark-25* is more challenging than that on *landmark-3D*. This is because landmark images in *landmark-3D* have less visual diversity than images in *landmark-25*. Since real landmark images are distributed with high visual diversity, we construct our own datasets *landmark-25* and *landmark-101* in this paper to evaluate the performance of the proposed approach on more real landmark images.

In addition, for $HMME + SVM$, the classification time for a given image on *Landmark-3D*, *Landmark-25* and *Landmark-101* is 35 ms, 35 ms, and 46 ms respectively. Thus, we can conclude that $HMME + SVM$ is very efficient landmark classification scheme.

B. Discussion

In this section, we provide comprehensive analysis to discuss factors that are most related to the performance of the proposed approach. Specifically, we explore effects of multi-modal feature fusion, effects of hierarchical partition, effects of exemplar selection, and effects of dimension reduction, on the overall performance. All the following experiments are conducted on dataset *Landmark-25*. Similar results can also be obtained on other two datasets.

1) Effects of Multi-Modal Feature Fusion: In our approach, discriminative information from multiple modalities are integrated into HMME. In fact, landmark images potentially contain large quantities of heterogeneous information from aspects of color, texture, shape, and appearance, which can be characterized by features extracted from the corresponding modalities.

TABLE II
CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT FEATURE CONFIGURATIONS

Feature configurations	Classification accuracy (%)
HCME	30±0.46
HLBPE	51.52±0.67
HHOGE	45.16±0.82
HBoVWE	55.64±0.53
HMME	**73.72±0.49**

TABLE III
PERFORMANCE LOSS WHEN DIFFERENT FEATURES ARE REMOVED FROM THE CONSTRUCTION PROCESS OF HMME

Feature configurations	Classification accuracy (%)
HMME\CM	72.92±0.75
HMME\LBP	69.8±0.61
HMME\HOG	71.6±0.56
HMME\BoVW	63.28±0.64
HMME	**73.72±0.49**

These features may make up their advantages and disadvantages, and integrating them may make new contributions on improving the performance. In this subsection, experiment is conducted to explore the above possibility on HMME based landmark classification. Performance of multi-modal feature fusion and that of single feature based HMME are evaluated. Also, performances of single features are observed to find which feature performs better on characterizing visual contents of landmark images.

Table II summarizes the classification results. Feature configuration denoted by "HXXE" means only single feature "XX" is used in HMME. For example, "HCME" denotes only CM is used. From the presented results, we can clearly find that HMME performs better than any other feature configurations. Classification accuracy increases from 55.64% with HBoVWE to 73.72% with HMME. There is nearly 18% performance improvement. Among single features, HCME achieves the worst performance. It has 43.72% lower classification accuracy than HMME. This is because landmark images are generally photographed under various light conditions, which makes images in a landmark category more visually diverse in terms of color distribution. Distinguished from many conventional classification approaches, where shape feature achieves the best performance, HBoVWE and HLBPE perform better than other features in task of landmark classification. More specifically, HBoVWE can gain 25% and 10% better classification accuracy compared with HCME and HHOGE respectively. This is because landmark is generally comprised of repetitive local structures, which can be better characterized by texture and appearance feature. From the above experimental results, we can draw a conclusion that combing features from heterogeneous visual modalities into HMME can bring further performance improvement on landmark classification.

Table III shows the main experimental results when different features are removed from the construction process of HMME. In these results, "HMME\XX" denotes feature "XX" is removed from HMME. For example, HMME\CM denotes CM feature is removed from HMME construction. In other words, HMME is constructed with LBP, HOG, and BoVW. From the table, we can easily observe that:

TABLE IV
CLASSIFICATION ACCURACIES ACHIEVED WHEN DIFFERENT SPATIAL LEVELS ARE EXPLOITED

Level combination	Classification accuracy (%)
Level 1	59.64±0.58
Level 2	66.2±0.81
Level 3	70.48±0.72
Level 1+Level 2	68.16±0.46
Level 1+Level 3	71.92±0.39
Level 2+Level 3	72.76±0.55
HMME	**73.72±0.49**

- any feature configurations with feature removing generate more or less performance loss, which reveals that the features employed in this paper all contribute to the final performance. The performance loss is ranged from 0.8% to nearly 10%;
- different feature configurations generate different performance loss. For example, HMME\BoVW brings the maximum performance loss (nearly 10%), while HMME\LBP brings the second performance loss (nearly 4%). HMME\CM and HMME\HOG all bring small performance loss. This phenomenon is caused because the performance loss is highly related to the discriminative ability of features. Feature with high discriminative ability brings much more performance loss if it is removed from the construction process of HMME, and vise versa.

2) Effects of Hierarchical Partition: In our approach, images are hierarchically partitioned into different size of regions with grids. The main objective of hierarchical image partition is to incorporate feature distribution at multiple spatial levels into HMME. In principle, on the one side, more spatial information may be included into feature representation by combing features extracted from multiple segmented regions, which may play positive impact on the performance. On the other side, many feature dimensions will be included into the final feature representation, which may bring undesirable and negative noises. Therefore, this experiment mainly addresses two essential questions, as follows.

- For task of landmark classification, does hierarchical partition produce positive effect on system performance?
- What is the level size that we can adopt to achieve the best performance on landmark dataset?

In this subsection, we conduct experiments to observe the performance achieved on different level size and the performance obtained by combing features from multiple spatial levels. Experimental results shown in Table IV clearly demonstrate that feature extracted at the level 3 achieves the best performance in terms of classification rate among spatial levels. Combing features from different spatial levels further brings performance improvement. HMME achieves 3.2% better performance compared with the performance achieved on level 3. We believe the performance improvement is due to the fact that features extracted from higher levels can capture finer visual distribution, while features extracted from lower level can perform well on describing macro visual contents. Fusing these complementary information can complement each other properly and produce positive effects on classification performance. In addition, we find that there is no further performance improvement when we increase the level size further (more than 3). Therefore, to

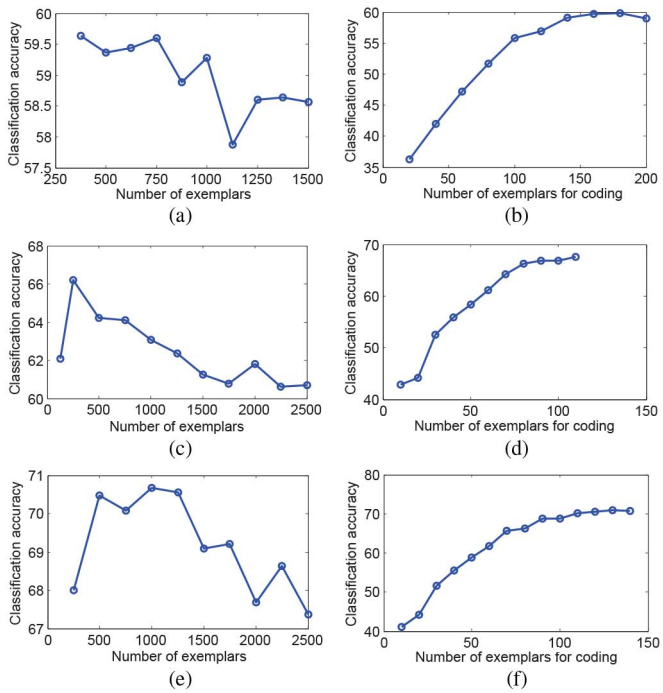

Fig. 5. Performance variations with parameters at different spatial levels. (a) Level 1. (b) Level 1. (c) Level 2. (d) Level 2. (e) Level 3. (f) Level 3.

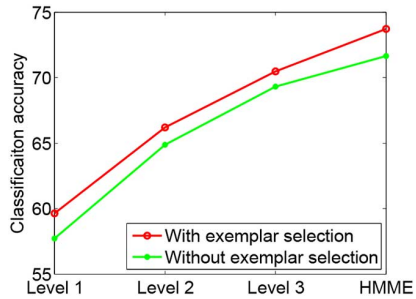

Fig. 6. Performance improvement at different spatial levels via exemplar selection.

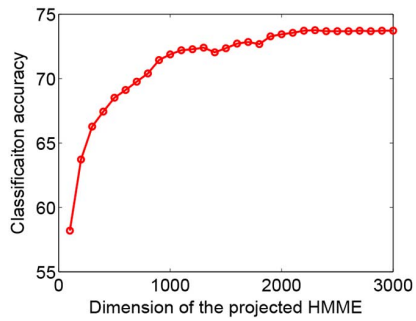

Fig. 7. Classification accuracy variations with the dimension of the projected HMME.

achieve high accuracy rate and meanwhile make the dimension of generated features as low as possible, the level size of hierarchical spatial partition is set to 3.

3) Effects of Exemplar Selection: With exemplar selection, hierarchical multi-modal exemplars, which are embedded with latent semantics of landmarks, can be discovered intelligently and automatically. Based on the discovered exemplars, optimal reconstruction coefficients are calculated between image instance and them, which construct the feature dimensions of HMME. Principally, the performance of exemplar selection depends on two important parameters: number of exemplars and number of exemplars for coding. In this experiment, we vary these two parameters to observe the performance variations at different spatial levels. Fig. 5 presents the main results. It demonstrates that accuracy curves in (a), (c), (e) all increase before a certain point, and decrease after that, while accuracy curves in (b), (d), (f) all increase steadily before a certain point, and they become steady after that. Therefore, we set the number of exemplars at three spatial levels as 250, 250, 500 respectively. The number of exemplars for coding at three levels are set to 160, 80, 120 respectively. In this case, the best feature dimensions of HMME on three spatial levels are 1000, 4000, and 18000 respectively. Therefore, the best size of HMME feature vectors on *Landmark-25* is 23000.

In our approach, multi-modal exemplars are either global or regional images. We claim that, via exemplar selection, more semantics can be embedded in the discovered exemplars, and thus the generated HMME can be more discriminative on distinguishing images. To validate our claim, we conduct experiments to compare the performance achieve by our approach and the approach which selects exemplars randomly without exemplar selection. Fig. 6 demonstrates performance obtained at different spatial levels before and after exemplar selection. It can be easily observed from the figure that, our approach based on exemplar selection consistently performs better than its competitor on all spatial levels. The presented results demonstrate the effectiveness of the proposed approach on discovering semantic exemplars.

4) Effects of Dimension Reduction: In our approach, dimension reduction is exploited to reduce redundant and noisy information in HMME. The experimental study evaluates the effects of dimension reduction. We vary the dimension of projected feature and observe the performance. Fig. 7 illustrates the classification accuracy variations with the dimension of the projected feature. It is shown that, when the dimension is higher than a certain value (> 2200), accuracy ratio becomes stable. Therefore, the best reduced size of HMME on *Landmark-25* after applying PCA can be set to 2200. This experimental phenomenon demonstrates that PCA performs well on redundancy removal with proper parameter setting. We can also find that, when the dimension is lower than 2200, the accuracy decreases steadily. This experimental results can be easily explained that compressing feature excessively cause discriminative information loss.

VIII. Conclusions

Effective landmark classification is fundamental for many georeferenced image search and analytics applications. One of the most challenging problems in landmark classification is how to model and characterize landmark image, which could have high visual diversity and complexity. In this paper, we present a novel image signature scheme called HMME to effectively

characterize landmark images. Also, an effective exemplar selection approach is proposed to mine hierarchical exemplars in multiple modalities from large amount of candidate images and regions. An effective feature generation framework based on region coding is developed to generate feature dimensions. Based on hierarchical multi-modal exemplars, HMME can effectively represent diverse visual contents of landmark images. Further, with region based semantic coding, HMME can integrate heterogeneous discriminative information from multiple modalities and various spatial levels, and enjoy superior robustness against visual variance of landmark images. Comparative experiments on real world landmark datasets demonstrate the effectiveness of HMME compared with several state-of-the-art techniques. Superior performance of HMME illustrates its greatest potentials for being applied to a wide range of real world visual landmark retrieval and mining applications.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive and helpful suggestions.

REFERENCES

[1] Z. Cheng, J. Ren, J. Shen, and H. Miao, "Building a large scale test collection for effective benchmarking of mobile landmark search," in *Proc. Int. Conf. Adv. Multimedia Modeling*, 2013, pp. 36–46.

[2] J. Shen, Z. Cheng, J. Shen, T. Mei, and X. Gao, "The evolution of research on multimedia travel guide search and recommender systems," in *Proc. Int. Conf. Adv. Multimedia Modeling*, , 2014, pp. 227–238.

[3] J. Ye *et al.*, "Dlmsearch: Diversified landmark search by photo," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 905–908.

[4] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie, "Content-based visual landmark search via multimodal hypergraph learning," *IEEE Trans. Cybern*, to be published.

[5] S. Jiang, X. Qiang, J. Shen, Y. Fu, and T. Mei, "Author topic model based collaborative filtering for personalized POI recommendation," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, Jun. 2015.

[6] T. Chen and K. Yap, "Discriminative BoW framework for mobile landmark recognition," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 695–706, May 2014.

[7] W. Min, C. Xu, M. Xu, X. Xiao, and B. Bao, "Mobile landmark search with 3D models," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 623–636, Apr. 2014.

[8] X. Xiao, C. Xu, J. Wang, and M. Xu, "Enhanced 3-D modeling for landmark image classification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pt. 2, pp. 1246–1258, Aug. 2012.

[9] W. Min, B. Bao, and C. Xu, "Multimodal spatio-temporal theme modeling for landmark analysis," *IEEE Multimedia Mag.*, vol. 21, no. 3, pp. 20–29, Jul.–Sep. 2014.

[10] T. Chen, K. Yap, and D. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 612–622, Apr. 2014.

[11] Y.-T. Zheng *et al.*, "Tour the world: Building a web-scale landmark recognition engine," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1085–1092.

[12] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.(CVPR)*, Jun. 2008, pp. 1–8.

[13] D. Chen *et al.*, "City-scale landmark identification on mobile devices," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 737–744.

[14] Q. Hao *et al.*, "3D visual phrases for landmark recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3594–3601.

[15] R. Raguram, C. Wu, J.-M. Frahm, and S. Lazebnik, "Modeling and recognition of landmark image collections using iconic scene graphs," *Int. J. Comput. Vis.*, vol. 95, no. 3, pp. 213–239, 2011.

[16] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes Paris look like Paris," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 101:1–101:9, 2012.

[17] T. Chen and K. Yap, "Context-aware discriminative vocabulary learning for mobile landmark recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 9, pp. 1611–1621, Sep. 2013.

[18] Y. Li, D. Crandall, and D. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep.–Oct. 2009, pp. 1957–1964.

[19] A. Bergamo, S. N. Sinha, and L. Torresani, "Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 763–770.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[21] C. Sun and K.-M. Lam, "Multiple-kernel, multiple-instance similarity features for efficient visual object detection," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3050–3061, Aug. 2013.

[22] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 89–96.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[24] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, vol. 2, pp. 1470–1477.

[25] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 75–84.

[26] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*. San Mateo, CA, USA: Morgan Kaufmann, 1996.

[27] R. Ji *et al.*, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, 2012.

[28] K.-H. Yap, Z. Li, D.-J. Zhang, and Z.-K. Ng, "Efficient mobile landmark recognition based on saliency-aware scalable vocabulary tree," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 1001–1004.

[29] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, Jun. 1998.

[30] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[31] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 41–48.

[32] F. Mindru, T. Tuytelaars, L. V. Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *Comput. Vis. Image Understanding*, vol. 94, pp. 3–27, 2004.

[33] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recog.*, vol. 29, no. 1, pp. 51–59, 1996.

[34] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[35] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 2, pp. 524–531.

[36] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[37] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[38] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, Mar. 2012.

[39] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Adv. Neural Inf. Process. Syst.*, pp. 561–568, 2003.

[40] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, 1967, vol. 1, pp. 281–297.

[41] J. Shlens, "A tutorial on principal component analysis," *CoRR*, vol. abs/1404.1100, 2014.

[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.